Sustainable Solutions to Intellectual Property and Ethical Complexities in Building a National Corpus

Robyn Rebollo¹, Michael Haugh², Simon Musgrave³, Xiaobin Shen⁴

¹Griffith University, Brisbane, Australia, r.rebollo@griffith.edu.au
² Griffith University, Brisbane, Australia, m.haugh@griffith.edu.au
³Monash University, Melbourne, Australia, Simon.Musgrave@arts.monash.edu.au
⁴Australian National Data Service, Melbourne, Australia, xiaobin.shen@ands.org.au

PROJECT HISTORY

The Australian National Corpus (AusNC) is an ongoing project to collate and provide access to a wide range of samples of Australian language for use in academic research. Building on earlier work collecting corpora in different disciplines, AusNC will bring existing and newly collected samples together in one place and provide tools to help researchers annotate, analyse and collaborate on this data.[1]

In 2010 the Australian National Corpus Inc., an independent, non-profit association which has been created to oversee the development of the corpus,[2] received funding from the Australian National Data Service. The main objective is to establish an AusNC solution that aggregates data from existing corpora residing at Australian universities, thereby providing a diverse and accurate representation of written and spoken languages in Australia. The goal of increasing the potential for re-use of linguistic research datasets— and thus enabling new research opportunities, is also paramount.

LEGAL AND ETHICAL DILEMMAS

Important progress has been made in addressing the issues more generally for data sharing and management in the context of eResearch, for example, the Legal Framework for eResearch project.[3] However, a myriad of complicated legal and ethical issues have arisen from the project as an outcome of making large amounts of language data from an array of sources available to other researchers and the public, where permissible. These include the different forms of language data, the clearance of data where participants are used, and the ownership of dynamic annotations created from the language data, which remains to be addressed. A sound legal policy for the AusNC will depend on many factors, including those highlighted by Borgman in Scholarship in the Digital Age: local data practices, differences in the intellectual property rights intrinsic to data sources, and the need to maintain confidentiality of human subjects.[4] The issue over copyright is relevant to both retrospective and new collections of data. Copyright may apply to many collections of spoken language data, with copyright belonging to the originator of the collection. In order to ensure language rich content can be freely contributed to and accessed from the AusNC, it is important to implement a copyright agreement that at a minimum, allows the means to use, adapt, publish, reproduce and communicate language rich content for academic and research purposes through the AusNC discovery service. The standard copyright agreement must also compliment the basic terms stipulated the AusNC end user agreement. Several of the terms included in the end user agreement will be covered in the eResearch Australasia presentation.

AUSTRALIAN NATIONAL CORPUS LEGAL APPROACH

The AusNC project conducted a Collections for Inclusion (http://www.ausnc.org.au/corpus_included.html) audit to ascertain accurate information about the diverse existing holdings intended for inclusion to the AusNC. The project quickly learned that several collections have unique copyright and privacy requirements, and trying to accomodate everyone’s specialised needs was not practical, nor feasible. The AusNC project set up a legal group comprising of steering committee members who were also collection custodians. It was decided the utilisation of Creative Commons licenses was not an acceptable option for the AusNC, because of the legal complexities of some of the collections. The legal group agreed that prospective collection owners of the AusNC should, at a minimum, allow the following usage rights to their collections: the means to use, adapt, publish, reproduce and communicate language rich content for academic and research purposes through the AusNC discovery service. Additional copyright requirements regarding general terms and warranties were documented during both scheduled meetings and email correspondence. Drafts for a general copyright and end user licence agreement were produced with the help of AusNC legal group members and inside legal counsel at Griffith University. The copyright licence drafting performed by Griffith University counsel was prepared for Griffith specific collections, with the intention of further modifications to be implemented by outside counsel hired by the AusNC project. With permission from Griffith, it is expected that outside counsel will review, modify and adapt the Griffith copyright draft licence agreement to create a standard copyright template to be used by any collection owner, institutional or individual, that wishes to contribute their collections to the AusNC. Practical suggestions for other projects that require legal protections for their data based discovery services, along with the sample AusNC licences, will be covered in the eResearch Australasia presentation.
ETHICAL CLEARANCE: THE MONASH EXPERIENCE

The Monash Corpus of English (MCE) was collected in the mid 1990s by staff of the Linguistics program at Monash University. The original ethics clearances for the data collection assumed that only those staff members would use data; subsequently additional clearance had to be sought for staff members not originally involved to have access to data. Given this background, the custodian of the MCE thought it necessary to seek a new clearance in order for the data to be exposed via the AusNC. To clarify this issue, the custodian met with the Manager (Research Ethics and Compliance). The Data Management Co-ordinator of the Monash University Library, also attended the meeting. The outcome of that meeting was that an application for new use of the data would be necessary, and that two issues in particular would have to be addressed by that application.

The first issue was the question of whether it was possible or feasible to contact the participants in the research to seek consent for the new use of the data. The dataset is formally de-identified; that is, no list has been maintained linking anonymous codes to names and other contact details. In principle therefore, there would seem to be no possibility of contacting the participants. However, the situation is a little more complicated in that the first question in one of the interview schedules had the participants give their names on tape. Obviously, this information will be deleted from audio and transcript if the material is exposed via AusNC, however, this means that, in conjunction with the codes used for data, a name and a school is known for each participant. The case, which has been made to the Ethics Committee (at time of writing the outcome of the application is unknown), is that this is not sufficient to make finding the participants feasible. The full reasoning behind this conclusion will be covered in the eResearch Australasia presentation.

This argument is made in association with a response to the second concern that had to be addressed: the extent to which the material might cause distress or embarrassment to the participants. In the meeting discussed above, the interview schedules used in data collection were considered and several items were identified as being most likely to have generated potentially sensitive material. A research assistant was then charged with examining the relevant transcripts to assess the nature of the responses. Based on this research, it was possible to include examples of the material in the application for new use of data and to make the case that the data was not of a kind that could reasonably be considered likely to cause distress or embarrassment if published. This position also is relevant to the first issue discussed: given that the material is not sensitive, the importance of seeking new consent is reduced. Although the application has not received a decision as yet, we are confident that permission will be granted for the MCE to form a part of AusNC.

THE AUSTRALIAN NATIONAL DATA SERVICE PERSPECTIVE

ANDS is promoting the Australian Research Data Commons (ARDC) to improve data usage and re-usage in Australian research, an important objective of the AusNC project. This also includes better access to reference data from the public sector. Unfortunately, licensing and access policy barriers will diminish the usefulness of the Australian Research Data Commons infrastructure. ANDS is therefore keen to promote and facilitate implementation of an appropriate information licensing framework in two contexts [5]:

- implementation at government agencies to clarify and optimize re-use of agency data by the research and innovations sector; and
- implementation at research institutions to clarify and optimise re-use of research sector data products (which are often a value-add on top of public sector)

In both instances ANDS is interested in working through the data re-use issues in very practical ways with regard to actual reference data collections of relevance to the research sector. ANDS is also interested in promoting the data management policy and procedures that necessarily accompany any licensing declaration [5]. Therefore, the work undertaken by the AusNC will inform these issues.

CONCLUSION

While a number of progressive licensing frameworks [Creative CommonsAU, Australian Governments Open Access and Licensing (AusGOAL) Framework] are available for copyright owners to use to allow others to share and reuse their material legally, the full legal version of the licenses are not simple to interpret, and not always applicable to implement with intrinsic federated discovery services, such as the AusNC. However our presentation will demonstrate the value in allocating time to utilise the appropriate support channels (Legal Department & Research Office – Ethics) within your institution for advice and direction.

REFERENCES

ABOUT THE AUTHOR(S)

Robyn Rebollo has over 16 years experience in information management, data management, online research, library technologies and vendor relations. Her authority in the information industry has been well established, having contributed to the design and development of e-solutions in the U.S. and Australasia. Her current role is Senior Project Manager, eResearch Services, Griffith University. She presently manages the Australian National Corpus Project, an initiative that aims to provide a federated solution for Australia's language data, enabling researchers and educators to have access to a wide range of multi-modal language data.

Michael Haugh is a Senior Lecturer in Linguistics/International English teaching a range of second and third year courses in pragmatics, conversation analysis, intercultural communication, and sociolinguistics. His research interests are closely aligned with his teaching, which emphasises the importance of research-based learning and applying linguistic theory to real world data. He is an Associate Editor for Journal of Pragmatics (Elsevier, Amsterdam), and on the Editorial Board of Intercultural Pragmatics (Mouton de Gruyter, Berlin). He has edited a number of volumes on the topics of politeness, face, intention, and corpus linguistics. He is also leading the establishment of the Australian National Corpus.

Simon Musgrave is a linguist at Monash University and Treasurer of AusNC Inc, with a strong interest in data management in language documentation. Over the last few years, Simon has sought to contribute his knowledge and experience from language documentation to the design and administration of the AusNC project.

Xiaobin Shen is a software lead and integration support at Australian National Data Service. He is particularly interested in eResearch related technology and data management, information visualization and human computer interaction. He is also the ANDS engaged research analyst for the Australian National Corpus project.